EXHIBIT A

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 10915–10919, November 1992
Biochemistry

Amino acid substitution matrices from protein blocks

(amino acid sequence/alignment algorithms/data base searching)

STEVEN HENIKOFF* AND JORJA G. HENIKOFF

Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Communicated by Walter Gilbert, August 28, 1992 (received for review July 13, 1992)

ABSTRACT **Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. The most widely used matrices are based on the Dayhoff model of evolutionary rates. Using a different approach, we have derived substitution matrices from about 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins. This led to marked improvements in alignments and in searches using queries from each of the groups.**

Among the most useful computer-based tools in modern biology are those that involve sequence alignments of proteins, since these alignments often provide important insights into gene and protein function. There are several different types of alignments: global alignments of pairs of proteins related by common ancestry throughout their lengths, local alignments involving related segments of proteins, multiple alignments of members of protein families, and alignments made during data base searches to detect homology. In each case, competing alignments are evaluated by using a scoring scheme for estimating similarity. Although several different scoring schemes have been proposed (1–6), the mutation data matrices of Dayhoff (1, 7–9) are generally considered the standard and are often the default in alignment and searching programs. In the Dayhoff model, substitution rates are derived from alignments of protein sequences that are at least 85% identical. However, the most common task involving substitution matrices is the detection of much more distant relationships, which are only inferred from substitution rates in the Dayhoff model. Therefore, we wondered whether a better approach might be to use alignments in which these relationships are explicitly represented. An incentive for investigating this possibility is that implementation of an improved matrix in numerous important applications requires only trivial effort.

METHODS

Deriving a Frequency Table from a Data Base of Blocks. Local alignments can be represented as ungapped blocks with each row a different protein segment and each column an aligned residue position. Previously, we described an automated system, PROTOMAT, for obtaining a set of blocks given a group of related proteins (10). This system was applied to a catalog of several hundred protein groups, yielding a data base of >2000 blocks. Consider a single block representing a conserved region of a protein family. For a new member of this family, we seek a set of scores for matches and mismatches that best favors a correct alignment with each of the other segments in the block relative to an incorrect alignment. For each column of the block, we first count the number of matches and mismatches of each type between the new sequence and every other sequence in the block. For example, if the residue of the new sequence that aligns with the first column of the first block is A and the column has 9 A residues and 1 S residue, then there are 9 AA matches and 1 AS mismatch. This procedure is repeated for all columns of all blocks with the summed results stored in a table. The new sequence is added to the group. For another new sequence, the same procedure is followed, summing these numbers with those already in the table. Notice that successive addition of each sequence to the group leads to a table consisting of counts of all possible amino acid pairs in a column. For example, in the column consisting of 9 A residues and 1 S residue, there are 8 + 7 + . . . 1 = 36 possible AA pairs, 9 AS or SA pairs, and no SS pairs. Counts of all possible pairs in each column of each block in the data base are summed. So, if a block has a width of w amino acids and a depth of s sequences, it contributes $ws(s - 1)/2$ amino acid pairs to the count [$(1 \times 10 \times 9)/2 = 45$ in the above example]. The result of this counting is a frequency table listing the number of times each of the 20 + 19 + . . . 1 = 210 different amino acid pairs occurs among the blocks. The table is used to calculate a matrix representing the odds ratio between these observed frequencies and those expected by chance.

Computing a Logarithm of Odds (Lod) Matrix. Let the total number of amino acid i, j pairs ($1 \le j \le i \le 20$) for each entry of the frequency table be f_{ij}. Then the observed probability of occurrence for each i, j pair is

$$q_{ij} = f_{ij} / \sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}.$$

For the column of 9 A residues and 1 S residue in the example, where $f_{AA} = 36$ and $f_{AS} = 9$, $q_{AA} = 36/45 = 0.8$ and $q_{AS} = 9/45 = 0.2$. Next we estimate the expected probability of occurrence for each i, j pair. It is assumed that the observed pair frequencies are those of the population. For the example, 36 pairs have A in both positions of the pair and 9 pairs have A at only one of the two positions, so that the expected probability of A in a pair is $[36 + (9/2)]/45 = 0.9$ and that of S is $(9/2)/45 = 0.1$. In general, the probability of occurrence of the ith amino acid in an i, j pair is

$$p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2.$$

The expected probability of occurrence e_{ij} for each i, j pair is then $p_i p_j$ for $i = j$ and $p_i p_j + p_j p_i = 2p_i p_j$ for $i \neq j$. In the example, the expected probability of AA is $0.9 \times 0.9 = 0.81$, that of AS + SA is $2 \times (0.9 \times 0.1) = 0.18$, and that of SS is $0.1 \times 0.1 = 0.01$. An odds ratio matrix is calculated where each entry is q_{ij}/e_{ij}. A lod ratio is then calculated in bit units as $s_{ij} = \log_2(q_{ij}/e_{ij})$. If the observed frequencies are as expected, $s_{ij} = 0$; if less than expected, $s_{ij} < 0$; if more than expected, $s_{ij} > 0$. Lod ratios are multiplied by a scaling factor of 2 and then rounded to the nearest integer value to produce

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: lod, logarithm of odds.

*To whom reprint requests should be addressed.

BLOSUM (blocks substitution matrix) matrices in half-bit units, comparable to matrices generated by the PAM (percent accepted mutation) program (11). For each substitution matrix, we calculated the average mutual information (12) per amino acid pair H (also called relative entropy), and the expected score E in bit units as

$$H = \sum_{i=1}^{20} \sum_{j=1}^{i} q_{ij} \times s_{ij}; \qquad E = \sum_{i=1}^{20} \sum_{j=1}^{i} p_i \times p_j \times s_{ij}.$$

Clustering Segments Within Blocks. To reduce multiple contributions to amino acid pair frequencies from the most closely related members of a family, sequences are clustered within blocks and each cluster is weighted as a single sequence in counting pairs (13). This is done by specifying a clustering percentage in which sequence segments that are identical for at least that percentage of amino acids are grouped together. For example, if the percentage is set at 80%, and sequence segment A is identical to sequence segment B at ≥80% of their aligned positions, then A and B are clustered and their contributions are averaged in calculating pair frequencies. If C is identical to either A or B at ≥80% of aligned positions, it is also clustered with them and the contributions of A, B, and C are averaged, even though C might not be identical to both A and B at ≥80% of aligned positions. In the above example, if 8 of the 9 sequences with A residues in the 9A–1S column are clustered, then the contribution of this column to the frequency table is equivalent to that of a 2A–1S column, which contributes 2 AS pairs. A consequence of clustering is that the contribution of closely related segments to the frequency table is reduced (or eliminated when an entire block is clustered, since this is equivalent to a single sequence in which no substitutions appear). For example, clustering at 62% reduces the number of blocks contributing to the table by 25%, with the remainder contributing 1.25 million pairs (including fractional pairs), whereas without clustering, >15 million pairs are counted (Fig. 1). In this way, varying the clustering percentage leads to a family of matrices. The matrix derived from a data base of blocks in which sequence segments that are identical at ≥80% of aligned residues are clustered is referred to as BLOSUM 80, and so forth. The BLOSUM program implements matrix construction. Frequency tables, matrices, and programs for UNIX and DOS machines are available over Internet by anonymous ftp (sparky.fhcrc.org).

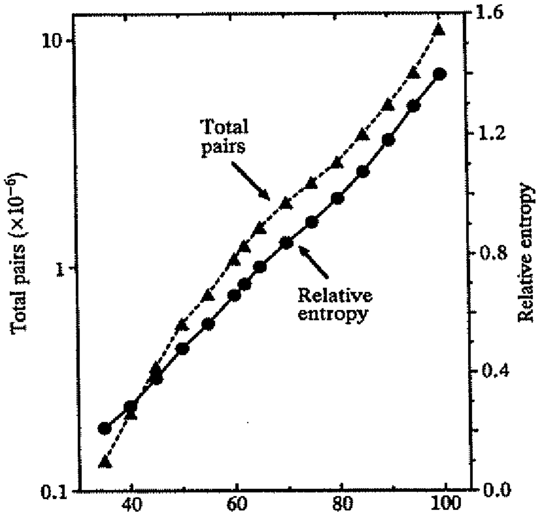


FIG. 1. Relationship between percentage clustering and total amino acid pair counts plotted on a logarithmic scale and relative entropy.

Constructing Blocks Data Bases. For this work, we began with versions of the blocks data base constructed by PROTOMAT (10) from 504 nonredundant groups of proteins catalogued in Prosite 8.0 (14) keyed to Swiss-Prot 20 (15). PROTOMAT employs an amino acid substitution matrix at two distinct phases of block construction (16). The MOTIF program uses a substitution matrix when individual sequences are aligned or realigned against sequence segments containing a candidate motif (16). The MOTOMAT program uses a substitution matrix when a block is extended to either side of the motif region and when scoring candidate blocks (10). A unitary substitution matrix (matches = 1; mismatches = 0) was used initially, generating 2205 blocks. Next, the BLOSUM program was applied to this data base of blocks, clustering at 60%, and the resulting matrix was used with PROTOMAT to construct a second data base consisting of 1961 blocks. The BLOSUM program was then applied to this second data base, clustering at 60%. This matrix was used to construct version 5.0 of the BLOCKS data base from 559 groups in Prosite 9.00 keyed to Swiss-Prot 22. The BLOSUM program was applied to this final data base of 2106 blocks, using a series of clustering percentages to obtain a family of lod substitution matrices. This series of matrices is very similar to the series derived from the second data base. Approximately similar matrices were also obtained from data bases generated by PROTOMAT using the PAM 120 matrix, using a matrix with a clustering percentage of 80%, and using just the odd- or even-numbered groups (data not shown).

Alignments and Homology Searches. Global multiple alignments were done using version 3.0 of MULTALIN for DOS computers (17). To provide a positive matrix, each entry was increased by 8 (with default gap penalty of 8). Version 1.6b2 of Pearson's RDF2 program (18) was used to evaluate local pairwise alignments.

Homology searches were done on a Sun Sparcstation using the BLASTP version of BLAST dated 3/18/91 (11) and version 1.6b2 of FASTA (with *ktup* = 1 and -o options) and SSEARCH, an implementation of the Smith–Waterman algorithm (18–20). The Swiss-Prot 20 data bank (15) containing 22,654 protein sequences was searched, and one search was done with each matrix for each of the 504 groups of proteins from Prosite 8.0. The first of the longest and most distant sequences in the group was chosen as a searching query, inferring distance from PROTOMAT results and Swiss-Prot names.

In the BLOSUM matrices, the scores for B and Z were made identical to those for D and E, respectively, and −1 was used for the character X. We used the same gap penalties for all matrices, −12 for the first residue in a gap, and −4 for subsequent residues in a gap.

The results of each search were analyzed by considering the sequences used by PROTOMAT to construct blocks for the protein group as the true positive sequences and all others as true negatives. BLAST reports the data bank matches up to a certain level of statistical significance. Therefore, we counted the number of misses as the number of true positive sequences not reported. For FASTA and SSEARCH, we followed the empirical evaluation criteria recommended by Pearson (19); the number of misses is the number of true positive scores, which ranked below the 99.5th percentile of the true negative scores.

RESULTS

Comparison to Dayhoff Matrices. The BLOSUM series derived from alignments in blocks is fundamentally different from the Dayhoff PAM series, which derives from the estimation of mutation rates. Nevertheless, the BLOSUM series based on percent clustering of aligned segments in blocks can be compared to the Dayhoff matrices based on PAM using a measure of average information per residue pair in bit units called relative entropy (9). Relative entropy is 0 when the target (or observed) distribution of pair frequencies is the same as the background (or expected) distribution and increases as these two distributions become more distinguishable. Relative entropy was used by Altschul (9) to characterize the Dayhoff matrices, which show a decrease with increasing PAM. For the BLOSUM series, relative entropy increases nearly linearly with increasing clustering percentage (Fig. 1). Based on relative entropy, the PAM 250 matrix is comparable to BLOSUM 45 with relative entropy of ≈0.4 bit, while PAM 120 is comparable to BLOSUM 80 with relative entropy of ≈1 bit. BLOSUM 62 (Fig. 2 *Lower*) is intermediate in both clustering percentage and relative entropy (0.7 bit) and is comparable to PAM 160. Matrices with comparable relative entropies also have similar expected scores.

Some consistent differences are seen when PAM 160 is subtracted from BLOSUM 62 for every matrix entry (Fig. 2 *Upper*). Compared to PAM 160, BLOSUM 62 is less tolerant to substitutions involving hydrophilic amino acids, while it is more tolerant to substitutions involving hydrophobic amino acids. For rare amino acids, especially cysteine and tryptophan, BLOSUM 62 is typically more tolerant to mismatches than is PAM 160.

Performance in Multiple Alignment of Known Structures. One test of sequence alignment accuracy is to compare the results obtained to alignments seen in three-dimensional structures. Lipman *et al.* (21) applied a simultaneous multiple alignment program, MSA, to 3 similarly diverged serine proteases of known three-dimensional structures. They found that for 161 closely aligned residue positions, 12 residues were involved in misalignments. We asked how well a hierarchical multiple alignment program, MULTALIN (17), performs on the same proteins using different substitution matrices. Table 1 shows that MULTALIN performs much worse than MSA using the PAM 120, 160, or 250 matrices, misaligning residues at 30–31 positions. In comparison, MULTALIN with a simple +6/−1 matrix (that assigns +6 to matches and −1 to mismatches) misaligns residues at 34 positions. In the same test using BLOSUM 45, 62 and 80, MULTALIN misaligned residues at only 6–9 positions. Comparable numbers were obtained when residues that show differences in the positions of side chains were excluded. Therefore, BLOSUM matrices produced accurate global alignments of these sequences.

Table 1. Performance of substitution matrices in aligning three serine proteases

Matrix aligned	Program	Residue positions missed*	
		All positions	Side chains
	MSA	12	6
PAM 120	MULTALIN	31	22
PAM 160	MULTALIN	30	22
PAM 250	MULTALIN	30	22
+6/−1	MULTALIN	34	26
BLOSUM 45	MULTALIN	9	5
BLOSUM 62	MULTALIN	6	4
BLOSUM 80	MULTALIN	9	6

*From data of Greer (22), where residues were considered to be aligned whenever α-carbons occupied comparable positions in space (All positions column). For a subset (Side chains column), residues were excluded where there were differences in the positions of side chains.

Performance in Searching for Homology in Sequence Data Banks. To determine how BLOSUM matrices perform in data bank searches, we first tested them on the guanine nucleotide-binding protein-coupled receptors, a particularly challenging group that has been used previously to test searching and alignment programs (10, 18, 23, 24). Three diverse queries, LSHRRAT, RTARAT, and UL33$HCMVA, were chosen from among the 114 full-length family members catalogued in Prosite based on the observation that none detected either of the others in searches. The number of misses was averaged in order to assess the overall searching performance of different matrices for this group. Three different programs were used—BLAST (11), FASTA (19), and Smith–Waterman (20). BLAST rapidly determines the best ungapped alignments in a data bank. FASTA is a heuristic and Smith–Waterman is a rigorous local alignment program; both can optimize an alignment by the introduction of gaps. Several BLOSUM and PAM matrices in the entropy range of 0.15–1.2 were tested.

Results with each of the 3 programs show that all BLOSUM matrices in the 0.3–0.8 range performed better than the best

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	−1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	−2	5	C
		2	0	−2	0	−1	0	0	0	1	0	0	0	1	0	1	−1	1	1	−1	S
C	9		2	−1	−1	−1	0	0	0	0	0	0	−1	0	−1	1	0	1	1	3	T
S	−1	4		2	−2	−1	−1	0	0	−1	−1	−1	1	1	0	−1	0	0	2	1	P
T	−1	1	5		2	−1	−2	−2	−1	0	0	1	1	0	0	1	0	1	1	2	A
P	−3	−1	−1	7		2	0	−1	−2	0	1	1	0	0	−1	0	−1	1	2	4	G
A	0	1	0	−1	4		3	−1	−1	0	0	1	−1	0	−1	0	−1	0	0	0	N
G	−3	0	−2	−2	0	6		2	−1	−1	−1	0	−1	0	0	0	0	2	1	3	D
N	−3	1	0	−2	−2	0	6		1	0	0	2	2	1	−1	0	0	2	2	4	E
D	−3	0	−1	−1	−2	−1	1	6		0	−2	0	1	1	−1	0	0	1	3	3	Q
E	−4	0	−1	−1	−1	−2	0	2	5		2	−1	0	1	0	−1	0	1	2	2	H
Q	−3	0	−1	−1	−1	−2	0	0	2	5		−1	−1	0	−1	1	0	1	3	−4	R
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8		1	−2	−1	1	1	2	3	1	K
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5		−2	−1	−1	0	1	2	4	M
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5		−1	1	0	0	1	3	I
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5		−1	0	−1	1	2	L
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4		0	1	2	4	V
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4		−1	−2	1	F
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4		−1	2	Y
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		−1	W
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7		
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

FIG. 2. BLOSUM 62 substitution matrix (*Lower*) and difference matrix (*Upper*) obtained by subtracting the PAM 160 matrix position by position. These matrices have identical relative entropies (0.70); the expected value of BLOSUM 62 is −0.52; that for PAM 160 is −0.57.

PAM matrix, PAM 200 (Fig. 3). In this range, each BLOSUM matrix missed 12–25 fewer members than the PAM matrix with similar relative entropy. Therefore, BLOSUM improved detection of members of this family regardless of the searching program used.

To determine whether the superiority of BLOSUM matrices over PAM matrices generalizes to other families, we carried out similar comparative tests for 504 groups of proteins catalogued in Prosite 8.0. For BLAST, BLOSUM 62 performed slightly better overall than BLOSUM 60 or 70, moderately better than BLOSUM 45, and much better than the best PAM matrix in this test, PAM 140 (Fig. 4). Specifically, BLOSUM 62 was better than PAM 140 for 90 groups, whereas it was worse in only 23 other groups. As a baseline for comparison, we used the simple +6/−1 matrix, which makes no distinction among matches or mismatches. Compared to +6/−1, BLOSUM 62 performance was better in 157 groups and was worse in 6 groups. Of the 504 groups tested, only 217 showed differences in any comparison. Similar results were obtained for FASTA (data not shown).

Very recently, two updates of the Dayhoff matrices have appeared (25, 26). Both use automated procedures to cluster similar sequences present within an entire protein data base and therefore provide considerably more aligned pairs than were used by Dayhoff. However, in tests of these matrices using BLAST on each of the 504 groups, performance was not noticeably different from that of the Dayhoff PAM 250 matrix, which these matrices were intended to replace, much worse than matrices in the BLOSUM series (Fig. 4). Compared to BLOSUM 45, which has similar relative entropy to PAM 250, the matrix of Gonnet *et al.* (25) was worse in 130 groups and better in only 3 groups and the matrix of Jones *et al.* (26) was worse in 138 groups and better in only 5 groups.

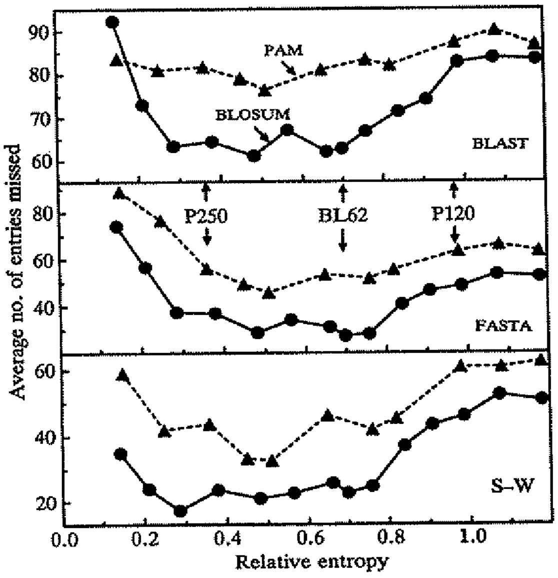


FIG. 3. Searching performance of programs using members of the guanine nucleotide-binding protein-coupled receptor family as queries and matrices from the BLOSUM and PAM series scaled in half-bits (11). Removal of this family from the BLOCKS data base led to a nearly identical matrix with similar performance. Matrices represented (left to right) are BLOSUM (BL) 30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, and 90 and PAM (P) 400, 310, 250, 220, 200, 160, 150, 140, 120, 110, and 100. The average numbers of true positive Swiss-Prot entries missed are shown for LSHRRAT, RTARAT, and UL33$HCMVA versus Swiss-Prot 20. Results using BLAST and FASTA or SSEARCH (S–W) are not comparable to each other, since different detection criteria were used for the three programs.

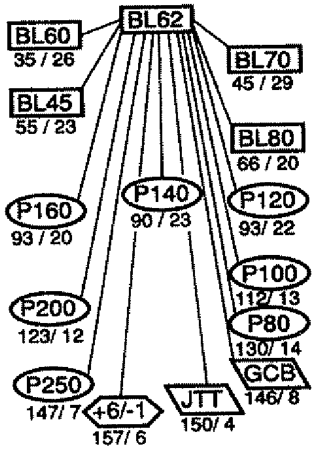


FIG. 4. Searching performance of BLAST using different matrices from the BLOSUM (BL) series, the PAM (P) series, and two recent updates of the standard Dayhoff matrix: GCB (25) and JTT (26). Results are based on searches using queries for each of 504 different groups. For each pair of numbers below a box representing a matrix, the first is the number of groups for which BLOSUM 62 missed fewer sequences than that matrix, and the second is the number of groups for which BLOSUM 62 missed more. The vertical distance between each matrix and BLOSUM 62 is proportional to the difference.

Confirmation of a Suspected Relationship Between Transposon Open Reading Frames. While the tests described above demonstrate that BLOSUM matrices perform better overall than PAM matrices, an example indicates the extent to which this improvement can matter in a real situation. We investigated a suspected relationship that is biologically attractive but is somewhat equivocal when examined by objective criteria. Two groups have noticed a stretch of similarity between the predicted protein from the *Drosophila mauritiana* mariner transposon and that from *Caenorhabditis elegans* transposon Tc1 (S. Emmons and J. Heierhorst, personal communications) (Fig. 5). However, this alignment did not score highly enough to allow its detection in searches using various PAM matrices. In contrast, a BLAST search with BLOSUM 62 using the mariner predicted protein as query detected this alignment as the best in the data base (data not shown). An analysis shows nonzero scores taken from the difference matrix of Fig. 2 assigned to each amino acid pair. The higher absolute score for BLOSUM 62 compared to PAM 160 ($\Sigma = 35$ for BLOSUM 62 > PAM 160 versus $\Sigma = 14$ for BLOSUM 62 < PAM 160) results from many small differences. When the scores for this alignment were compared to the scores for alignments between one of the sequences and 1000 shuffles of the other, the score using BLOSUM 62 was 7.6 SD above the mean. In contrast, the score using PAM 160 was only 3.0 SD above the mean with similar results for PAM 250 and PAM 120, accounting for the failure to detect this relationship in previous data base searches.

```
Mariner    IFLHDNAPSHTARAVRDTLETLNWEVLPHAAYSPDLAPSDY
            :  :: : ::   ::          :     :::: :
Tc1        VFQQDNDPKHTSLHVRSWFQRRHVHLLDWPSQSPDLNPIEH
BL62>P160      23 2 22 1    3   1 4    2  3222  2  2
BL62<P160   1 2  2        1  1       1  2     12  1
```

FIG. 5. Alignment of *D. mauritiana* mariner predicted protein (amino acids 245–295) with *C. elegans* TcA (amino acids 235–285) encoded by Tc1. Difference scores taken from Fig. 2 are indicated just below each alignment position. Using RDF2 with BLOSUM 62 for 1000 shuffles and a window size of 10, this alignment scores 64, compared to a mean of 31.4 (SD = 4.32) for $z = 7.6$. With PAM 160, the score is 43, compared to a mean of 30.1, SD = 4.63, and $z = 3.0$. With PAM 250, $z = 2.14$; with PAM 120, $z = 2.98$.

DISCUSSION

We have found that substitution matrices based on amino acid pairs in blocks of aligned protein segments perform better in alignments and homology searches than those based on accepted mutations in closely related groups. Performance was improved overall in every test we have done, including multiple alignment (MULTALIN), detection of ungapped alignments (BLAST), detection of gapped alignments (FASTA and Smith–Waterman), and determination of the significance of an alignment (RDF2). The importance of such improved performance can be profound for weakly scoring alignments that are not detected in a search or are not trusted. For example, the alignment between predicted proteins encoded by mariner and Tc1 transposons improved by more than 4.5 SD above the mean of comparisons to shuffled sequences when BLOSUM 62 was used instead of PAM matrices.

There are fundamental differences between our approach and that of Dayhoff that could account for the superior performance of BLOSUM matrices in searches and alignments. Dayhoff estimated mutation rates from substitutions observed in closely related proteins and extrapolated those rates to model distant relationships. In our case, frequencies were obtained directly from relationships represented in the blocks, regardless of evolutionary distance. Since blocks were derived primarily from the most highly conserved regions of proteins, it is possible that many of the differences between BLOSUM and PAM matrices arise from different constraints on conserved regions in general. For example, Dayhoff found asparagine to be the most mutable residue, whereas, in blocks, asparagine is involved in substitutions at an average frequency. This could mean that an asparagine located in a mutable region of a protein is itself highly mutable, whereas, when it is located in a conserved region, it shows only an average tendency to be involved in substitutions.

Another difference is the larger and more representative data set used in this work. The Dayhoff frequency table included 36 pairs in which no accepted point mutations occurred. In contrast, the pairs we counted included no fewer than 2369 occurrences of any particular substitution. Scoring differences were especially apparent for pairs involving rare amino acids such as tryptophan and cysteine. Similar findings were made in the two recent updates of the Dayhoff matrix (25, 26). However, in these studies, no evidence was presented that increased data improved performance. Our tests show that the updated Dayhoff matrices still perform poorly overall when compared to BLOSUM 62. This suggests that matrices from aligned segments in blocks, which represent the most highly conserved regions in proteins, are more appropriate for searches and alignments than are matrices derived by extrapolation from mutation rates.

The BLOSUM series depends only on the identity and composition of groups in Prosite and the accuracy of the automated PROTOMAT system. While the system itself uses a substitution matrix, iterative application soon leads to nearly the same set of scores, even starting with a unitary matrix or using a representative subset of the groups. Therefore, we do not expect that these substitution matrices will change significantly in the future.

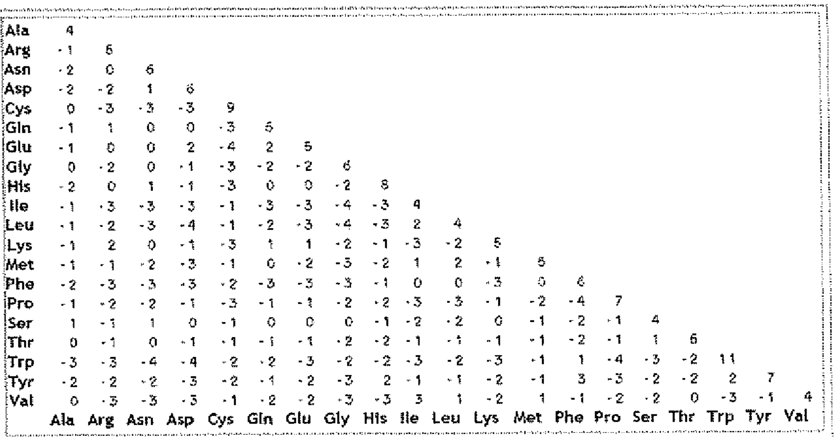
The suggestion to make a substitution matrix from a blocks data base was made by Temple Smith at the 1991 Aspen Center for Physics workshop. We thank Scott Emmons and Jörg Heierhorst for independently pointing out the similarity between mariner and Tc1 predicted proteins, Bill Pearson for advice, and Domokos Vermes for discussions about information theory. This work was supported by a grant from the National Institutes of Health.

1. Dayhoff, M. O. & Eck, R. V., eds. (1968) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 3, p. 33.
2. McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424.
3. Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985) *J. Mol. Evol.* **21**, 112–125.
4. Rao, J. K. M. (1987) *Int. J. Pept. Protein Res.* **29**, 276–281.
5. Risler, J. L., Delorme, M. O., Delacroix, H. & Henaut, A. (1988) *J. Mol. Biol.* **204**, 1019–1029.
6. Smith, R. F. & Smith, T. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 118–122.
7. George, D. G., Barker, W. C. & Hunt, L. T. (1990) *Methods Enzymol.* **183**, 333–351.
8. Dayhoff, M. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345–358.
9. Altschul, S. F. (1991) *J. Mol. Biol.* **219**, 555–565.
10. Henikoff, S. & Henikoff, J. G. (1991) *Nucleic Acids Res.* **19**, 6565–6572.
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
12. Blahut, R. E. (1987) *Principles and Practice of Information Theory* (Addison–Wesley, Reading, MA).
13. Henikoff, S., Wallace, J. C. & Brown, J. P. (1990) *Methods Enzymol.* **183**, 111–132.
14. Bairoch, A. (1991) *Nucleic Acids Res.* **19**, 2241–2245.
15. Bairoch, A. & Boeckmann, C. (1991) *Nucleic Acids Res.* **19**, 2247–2249.
16. Smith, H. O., Annau, T. M. & Chandrasegaran, S. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 826–830.
17. Corpet, F. (1988) *Nucleic Acids Res.* **16**, 10881–10890.
18. Pearson, W. R. (1990) *Methods Enzymol.* **183**, 63–98.
19. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
20. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
21. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4412–4415.
22. Greer, J. (1981) *J. Mol. Biol.* **153**, 1027–1042.
23. Doolittle, R. F. (1990) *Methods Enzymol.* **183**, 99–110.
24. Attwood, T. K., Eliopoulos, E. E. & Findlay, J. B. C. (1991) *Gene* **98**, 153–159.
25. Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992) *Science* **256**, 1443–1445.
26. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comp. Appl. Biosci.* **8**, 275–282.

BLOSUM

From Wikipedia, the free encyclopedia

BLOSUM (BLOcks of Amino Acid SUbstitution Matrix[1]) is a substitution matrix used for sequence alignment of proteins. BLOSUM are used to score alignments between evolutionarily divergent protein sequences. BLOSUM is based on local alignments. BLOSUM was first introduced in a paper by Henikoff and Henikoff.[2] They scanned the BLOCKS database for very conserved regions of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of amino acids and their substitution probabilities. Then, they calculated a log-odds score for each of the 210 possible substitutions of the 20 standard amino acids. All BLOSUM are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.



Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

The BLOSUM62 matrix

Several sets of BLOSUM exist using different alignment databases, named with numbers. BLOSUM with high numbers are designed for comparing closely related sequences, while BLOSUM with low numbers are designed for comparing distant related sequences. For example, BLOSUM80 is used for less divergent alignments, and BLOSUM45 is used for more divergent alignments. The matrices were created by merging (clustering) all sequences that were more similar than a given percentage into one single sequence and then comparing those sequences (that were all more divergent than the given percentage value) only; thus reducing the contribution of closely related sequences. The percentage used was appended to the name, giving BLOSUM80 for example where sequences that were more than 80% identical were clustered.

Scores within a BLOSUM are log-odds scores that measure, in an alignment, the logarithm for the ratio of the likelihood of two amino acids appearing with a biological sense and the likelihood of the same amino acids appearing by chance.[3] The matrices are based on the minimum percentage identity of the aligned protein sequence used in calculating them.[3] Every possible identity or substitution is assigned a score based on its observed frequences in the alignment of related proteins.[4] A positive score is given to the more likely substitutions while a negative score is given to the less likely substitutions.

To calculate a matrix for BLOSUM, the following equation is used: $S_{ij} = \left(\frac{1}{\lambda}\right) \log \left(\frac{p_{ij}}{q_i * q_j}\right)$

Here, p_{ij} is the probability of two amino acids i and j replacing each other in a homologous sequence, and q_i and q_j are the background probabilities of finding the amino acids i and j in any protein sequence at random. The factor λ is a scaling factor, set such that the matrix contains easily computable integer values.

References

1. ^ Note that in the acronym BLOSUM the last 'M' stands for 'matrix' and it is therefore incorrect and unnecessary to write 'BLOSUM matrix', see RAS syndrome.
2. ^ Henikoff, S. (1992). "Amino Acid Substitution Matrices from Protein Blocks". *PNAS* **89**: 10915–10919. doi:10.1073/pnas.89.22.10915. PMID 1438297. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=EBI&pubmedid=1438297.
3. ^ *a b* Albert Y. Zomaya (2006). *Handbook of Nature-Inspired And Innovative Computing.* New York, NY: Springer. ISBN 0387405321. http://books.google.com/books?id=kDFltuQo1dMC&pg=PA673&lpg=PA673&dq=blosum+matrix&source=web&ots=LBo5qtEF60&sig=o-

cOPVGPT_HPaPuw_EgTHTPRIEc#PPA697,M1.page 673

4. ^ NIH "Scoring Systems"

External links

- Page on BLOSUM
- Sean R. Eddy (2004). "Where did the BLOSUM62 alignment score matrix come from?". *Nature Biotechnology* **22**: 1035. doi:10.1038/nbt0804-1035. PMID 15286655. http://informatics.umdnj.edu/bioinformatics/courses/5020/notes/BLOSUM62%20primer.pdf.
- BLOCKS WWW server
- Scoring systems for BLAST at NCBI
- Data files of BLOSUM on the NCBI FTP server.

See also

- Sequence alignment
- Point accepted mutation

Retrieved from "http://en.wikipedia.org/wiki/BLOSUM"
Categories: Genetics | Bioinformatics | Biochemistry methods | Evolution | Computational phylogenetics | Matrices

- This page was last modified on 2 April 2009, at 21:21 (UTC).

Glycosylated

Glycosylated haemoglobin

Glycosylated hemoglobin

EXHIBIT B

Glycosylation

From Wikipedia, the free encyclopedia

Glycosylation is the enzymatic process that links saccharides to produce glycans, either free or attached to proteins and lipids. This enzymatic process produces one of four fundamental components of all cells (along with nucleic acids, proteins, and lipids) and also provides a co-translational and post-translational modification mechanism that modulates the structure and function of membrane and secreted proteins. The majority of proteins synthesized in the rough ER undergo glycosylation. It is an enzyme-directed site-specific process, as opposed to the non-enzymatic chemical reaction of glycation. Glycosylation is also present in the cytoplasm and nucleus as the O-GlcNAc modification. Six classes of glycans are produced: *N*-linked glycans attached to the amide nitrogen of asparagine side chains, *O*-linked glycans attached to the hydroxy oxygen of serine and threonine side chains; glycosaminoglycans attached to the hydroxy oxygen of serine; glycolipids in which the glycans are attached to ceramide, hyaluronan which is unattached to either protein or lipid, and GPI anchors which link proteins to lipids through glycan linkages.

Contents

Purpose

The polysaccharide chains attached to the target proteins serve various functions. For instance, some proteins do not fold correctly unless they are glycosylated first. Also, polysaccharides linked at the amide nitrogen of asparagine in the protein confer stability on some secreted glycoproteins. Experiments have shown that glycosylation in this case is not a strict requirement for proper folding, but the unglycosylated protein degrades quickly. Glycosylation may play a role in cell-cell adhesion (a mechanism employed by cells of the immune system), as well.

Mechanisms

There are various mechanisms for glycosylation, although all share several common features:

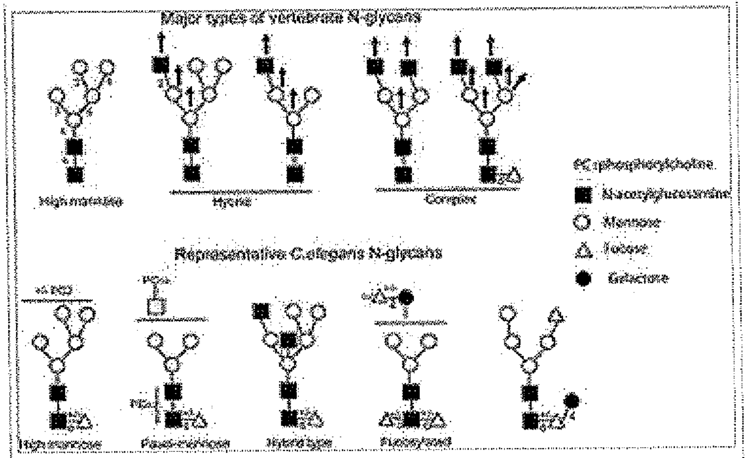
- Glycosylation is an enzymatic process;
- The donor molecule is an activated nucleotide sugar;
- The process is site-specific.

N-linked glycosylation

N-linked glycosylation is important for the folding of some eukaryotic proteins. The *N*-linked glycosylation process occurs in eukaryotes and widely in archaea, but very rarely in bacteria.

For *N*-linked oligosaccharides, a 14-sugar precursor is first added to the asparagine in the polypeptide chain of the target protein. The structure of this precursor is common to most eukaryotes, and contains 3 glucose, 9 mannose, and 2 *N*-acetylglucosamine molecules. A complex set of reactions attaches this branched chain to a carrier molecule called dolichol, and then it is transferred to the appropriate point on the polypeptide chain as it is translocated into the ER lumen.

There are three major types of *N*-linked saccharides: high-mannose oligosaccharides, complex oligosaccharides and hybrid oligosaccharides.



Comparative overview of the major types of vertebrate N-glycan subtypes and some representative *C. elegans* N-glycans.

- High-mannose is, in essence, just two *N*-acetylglucosamines with many mannose residues, often almost as many as are seen in the precursor oligosaccharides before it is attached to the protein.

- Complex oligosaccharides are so named because they can contain almost any number of the other types of saccharides, including more than the original two *N*-acetylglucosamines.

Proteins can be glycosylated by both types of oligos on different portions of the protein. Whether an oligosaccharide is high-mannose or complex is thought to depend on its accessibility to saccharide-modifying proteins in the Golgi. If the saccharide is relatively inaccessible, it will most likely stay in its original high-mannose form. If it is accessible, then it is likely that many of the mannose residues will be cleaved off and the saccharide will be further modified by the addition of other types of group as discussed above.

The oligosaccharide chain is attached by oligosaccharyltransferase to asparagine occurring in the tripeptide sequence Asn-X-Ser or Asn-X-Thr where X could be any amino acid except Pro. This sequence is known as a glycosylation *sequon*. After attachment, once the protein is correctly folded, the three glucose residues are removed from the chain and the protein is available for export from the ER. The glycoprotein thus formed is then transported to the Golgi where removal of further mannose residues may take place. However, glycosylation itself does not seem to be as necessary for correct transport targeting of the protein, as one might think. Studies involving drugs that block certain steps in glycosylation, or mutant cells deficient in a glycosylation enzyme, still produce otherwise-structurally-normal proteins that are correctly targeted, and this interference does not seem to interfere severely with the viability of the cells. Mature glycoproteins may contain a variety of oligomannose *N*-linked oligosaccharides containing between 5 and 9 mannose residues. Further removal of mannose residues leads to a 'core' structure containing 3 mannose, and 2 *N*-acetylglucosamine residues, which may then be elongated with a variety of different monosaccharides including galactose, *N*-acetylglucosamine, *N*-acetylgalactosamine, fucose and sialic acid.

O-linked glycosylation

O-N-acetylgalactosamine (*O*-GalNAc)

O-linked glycosylation occurs at a later stage during protein processing, probably in the Golgi apparatus. This is the addition of N-acetyl-galactosamine to serine or threonine residues by the enzyme *UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase* (EC 2.4.1.41), followed by other carbohydrates (such as galactose and sialic acid). This process is important for certain types of proteins such as proteoglycans, which involves the addition of glycosaminoglycan chains to an initially unglycosylated "proteoglycan core protein." These additions are usually serine *O*-linked glycoproteins, which seem to have one of two main functions. One function involves secretion to form components of the extracellular matrix, adhering one cell to another by interactions between the large sugar complexes of proteoglycans. The other main function is to act as a component of mucosal secretions, and it is the high concentration of carbohydrates that tends to give mucus its "slimy" feel. Proteins that circulate in the blood are not normally *O*-glycosylated, with the exception of IgA1 and IgD (two types of antibody) and C1-inhibitor.

O-fucose

O-fucose is added between the second and third conserved cysteines of EGF-like repeats in the Notch protein, and other substrates by GDP-fucose protein *O*-fucosyltransferase 1, and to Thrombospondin repeats by GDP-fucose protein *O*-fucosyltransferase 2. In the case of EGF-like repeats, the *O*-fucose may be further elongated to a tetrasaccharide by sequential addition of N-acetylglucosamine (GlcNAc), galactose, and sialic acid, and for Thrombospondin repeats, may be elongated to a disaccharide by the addition of glucose. Both of these fucosyltransferases have been localized to the endoplasmic reticulum, which is unusual for glycosyltransferases, most of which function in the Golgi apparatus.

O-glucose

O-glucose is added between the first and second conserved cysteines of EGF-like repeats in the Notch protein, and possibly other substrates by an unidentified *O*-glucosyltransferase.

O-*N*-acetylglucosamine (*O*-GlcNAc)

O-GlcNAc is added to serines or threonines by *O*-GlcNAc transferase. *O*-GlcNAc appears to occur on serines and threonines that would otherwise be phosphorylated by serine/threonine kinases. Thus, if phosphorylation occurs, *O*-GlcNAc does not, and *vice versa*. This is an incredibly important finding because phosphorylation/dephosphorylation has become a scientific paradigm for the regulation of signaling within cells. A massive amount of cancer research is focused on phosphorylation. Ignoring the involvement of this form of glycosylation, which clearly appears to act in concert with phosphorylation, means that a lot of current research is missing at least half of the picture. *O*-GlcNAc addition and removal also appear to be key regulators of the pathways that are deregulated in diabetes mellitus. The gene encoding the *O*-GlcNAc removal enzyme has been linked to non-insulin dependent diabetes mellitus. It is the terminal step in a nutrient-sensing hexosamine signaling pathway.

GPI anchor

A special form of glycosylation is the *GPI anchor*. This form of glycosylation functions to attach a protein to a hydrophobic lipid anchor, via a glycan chain. (see also prenylation)

C-mannosylation

A mannose sugar is added to tryptophan residues in Thrombospondin repeats. This is an unusual modification both because the sugar is linked to a carbon rather than a reactive atom like a nitrogen or oxygen and because the sugar is linked to a tryptophan residue rather than an asparagine or serine/threonine.

See also

- Glycation
- Advanced glycation endproduct
- Chemical glycosylation

External links

- Online textbook of glycobiology with chapters about glycosylation

- GlyProt: In-silico N-glycosylation of proteins on the web
- NetNGlyc: The NetNglyc server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons.

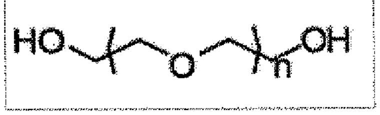
Retrieved from "http://en.wikipedia.org/wiki/Glycosylation"
Categories: Posttranslational modification | Organic chemistry | Carbohydrates | Carbohydrate chemistry
Hidden categories: Articles lacking sources from December 2007 | All articles lacking sources

- This page was last modified on 24 April 2009, at 13:41 (UTC).

EXHIBIT C

PEGylation

From Wikipedia, the free encyclopedia



Polyethylene glycol

PEGylation is the process of covalent attachment of poly(ethylene glycol) polymer chains to another molecule, normally a drug or therapeutic protein. PEGylation is routinely achieved by incubation of a reactive derivative of PEG with the target macromolecule. The covalent attachment of PEG to a drug or therapeutic protein can "mask" the agent from the host's immune system (reduced immunogenicity and antigenicity), increase the hydrodynamic size (size in solution) of the agent which prolongs its circulatory time by reducing renal clearance. PEGylation can also provide water solubility to hydrophobic drugs and proteins.

Contents

History

In 1970s, pioneering research by Dr. Frank Davis, Dr. Abraham Abuchowski and colleagues foresaw the potential of the conjugation of Polyethylene glycol (PEG) to proteins. Dr. Abuchowski founded Enzon, Inc, which brought three PEGylated drugs to market, and is the founder and president of Prolong Pharmaceuticals.

PEGylation, is a process of attaching the strands of the polymer PEG to molecules most typically peptides, proteins, and antibody fragments, that can help to meet the challenges of improving the safety and efficiency of many therapeutics. It produces alterations in the physiochemical properties including changes in conformation, electrostatic binding, hydrophobicity etc. These physical and chemical changes increase systemic retention of the therapeutic agent. Also, it can influence the binding affinity of the therapeutic moiety to the cell receptors and can alter the absorption and distribution patterns.

PEGylation, by increasing the molecular weight of a molecule, can impart several significant pharmacological advantages over the unmodified form, such as:

- Improved drug solubility

- Reduced dosage frequency, without diminished efficacy with potentially reduced toxicity

- Extended circulating life

- Increased drug stability

- Enhanced protection from proteolytic degradation

The PEGylated drugs are having the following commercial advantages also:

- Opportunities for new delivery formats and dosing regimens

- Extended patent life of previously approved drugs

PEGylated Pharmaceuticals on the Market

The clinical value of PEGylation is now well established. ADAGEN (PEG- bovine adenosine deaminase) manufactured by Enzon Pharmaceuticals, Inc., US was the first PEGylated protein approved by FDA in March 1990, to enter the market. It is used to treat X-linked severe combined immunogenicity syndrome, as an alternative to bone marrow transplantation and enzyme replacement by gene therapy. Since the introduction of ADAGEN, a large number of PEGylated protein and peptide pharmaceuticals have followed and many others are under clinical trial or under development stages. Some of the successful examples are:

- **PEGASYS:** PEGylated interferon alpha for use in the treatment of chronic hepatitis C and hepatitis B (Hoffman-La Roche)

- **Pegintron:** PEGylated interferon alpha for use in the treatment of chronic hepatitis C and hepatitis B (Schering-Plough / Enzon)

- **Oncaspar:** PEGylated L-asparaginase for the treatment of acute lymphoblastic leukemia in patients who are hypersensitive to the native unmodified form of L-asparaginase (Enzon). This drug was recently approved for front line use.

- **Neulasta:** PEGylated recombinant methionyl human granulocyte colony-stimulating factor for severe cancer chemotherapy induced neutropenia (Amgen)

- **Doxil:** PEGylated liposome containing doxorubicin for the treatment of Cancer (Sequus)

PEG Moiety Properties

PEG is a particularly attractive polymer for conjugation. The specific characteristics of PEG moieties relevant to pharmaceutical applications are:

- Water solubility

- High mobility in solution

- Lack of toxicity and immunogenicity

- Ready clearance from the body

- Altered distribution in the body

PEGylation Process

The first step of the PEGylation is the suitable functionalization of the PEG polymer at one or both terminals. PEGs that are activated at each terminus with the same reactive moiety are known as "homobifunctional", where as if the functional groups present are different, then the PEG derivative is referred as "heterobifunctional" or "heterofunctional." The chemically active or activated derivatives of the PEG polymer are prepared to attach the PEG to the desired molecule.

The choice of the suitable functional group for the PEG derivative is based on the type of available reactive group on the molecule that will be coupled to the PEG. For proteins, typical reactive amino acids include lysine, cysteine, histidine, arginine, aspartic acid, glutamic acid, serine, threonine, tyrosine. The N-terminal amino group and the C-terminal carboxylic acid can also be used.

The techniques used to form first generation PEG derivatives are generally reacting the PEG polymer with a group that is reactive with hydroxyl groups, typically anhydrides, acid chlorides, chloroformates and carbonates. In the second generation PEGylation chemistry more efficient functional groups such as aldehyde, esters, amides etc made available for conjugation.

As applications of PEGylation have become more and more advanced and sophisticated, there has been an increase in need for heterobifunctional PEGs for conjugation. These heterobifunctional PEGs are very useful in linking two entities, where a hydrophilic, flexible and biocompatible spacer is needed. Preferred end groups for heterobifunctional PEGs are maleimide, vinyl sulfones, pyridyl disulfide, amine, carboxylic acids and NHS esters.

References

Abuchowski, McCoy, Palczuk, van Es and Davis (1977). "Effect of covalent attachment of polyethylene glycol on immunogenicity and circulating life of bovine liver catalase." Journal of Biological Chemistry 252(11): 3582-3586.

Fee, Conan (2003). "Size exclusion reaction chromatography (SERC): A new technique for protein PEGylation." Biotechnology and Bioengineering 82(2): 200-206.

Fee, Conan and Van Alstine, J. M. (2006). "PEG-proteins: Reaction engineering and separation issues." Chemical Engineering Science 61(3): 924-939.

Fee, Conan (2007). "Size comparison between proteins PEGylated with branched and linear poly(ethylene glycol) molecules". Biotechnology and Bioengineering 98(4): 725-31.

Kodera, Matsushima, Hiroto, Nishimura, Ishii, Ueno and Inada (1998). "Pegylation of proteins and bioactive substances for medical and technical applications." Progress in Polymer Science 23(7): 1233-1271.

Morar, Jeffrey and Mark (2006). "PEGylation of Proteins: A Structural Approach." Biopharm International 19 (4): 34.

Roberts, Bentley and Harris (2002). "Chemistry for peptide and protein PEGylation." Advanced Drug Delivery Reviews 54(4): 459-476.

Veronese (2001). "Peptide and protein PEGylation: a review of problems and solutions." Biomaterials 22(5): 405-417.

Veronese and Harris (2002). "Introduction and overview of peptide and protein pegylation." Advanced Drug Delivery Reviews 54(4): 453-456.

Veronese and Pasut (2005). "PEGylation, successful approach to drug delivery." Drug Discovery Today 10 (21): 1451-1458.

See also

- Interferon
- Polyethylene glycol
- Size exclusion chromatography
- Proteomics
- Matrix-assisted laser desorption/ionization
- Granulocyte colony-stimulating factor
- Cytochrome c

Retrieved from "http://en.wikipedia.org/wiki/PEGylation"
Categories: Organic reactions
Hidden categories: Cleanup from June 2008 | Articles with sections that need to be turned into prose | Articles lacking in-text citations from April 2009